**Documentation**

**Bytelearn NLP Challenge**

**Challenges faced the given problem:**

1. The data provided is in the json format. To make it usable and python-ready, I had to convert into a data-frame.
2. It was difficult to choose between NLTK and Spacy as both of them have their advantage and disadvantages. So, I decided to work with both of these and extracted required entities.

* NLTK is better suited for tokenization and has support for wide number of languages although in the given challenge it didn’t matter as the dataset is in English.
* For every string input NLTK outputs string whereas, Spacy outputs a object.
* The most important thing is that Spacy has support for word vectors which NLTK lacks.

1. The next challenge was to extract only numerical values as those are the most important things while dealing with a mathematical equation. For that I had to filter the entities accordingly and had to match them with labels to create a dictionary of variables and coefficients. Which can further be used to devise an actual mathematical equation.
2. I tried formulating the word problem by using the entity values that my algorithm gave me. Albeit, I found it rather difficult since we use different kinds of language structures to say the same thing.
   1. For instance.: Consider this example: “Mary had 4 apples. Her father gave her 6 more apples. How many apples does Mary have now? “
   2. It is quite obvious to use that this problem is an addition problem and we can tell so from the word “gave”. Although, if we train our model to do addition every time it says “gave” it will be a chaos as follows:
   3. “Mary had 4 apples. She gave 2 apples to her friend How many apples does she have now?”
   4. Here, the problem has similar kind of structure and words used. Although, the word “gave” here tells us to subtract the values.
   5. It was really challenging for me to develop such a model in a day. But, given a chance I would like to work on such scenario.
3. In order to deploy a Flask app, I had to use Heroku. I tried pythonanywhere although it wasn’t successful.
4. For getting response using curl command was also challenging, as I had never passed form data using it.
5. I performed the EDA part in Google Colab in order to identify the pattern in the dataset and found that there are 4.125 entities in every sample. 60% of these total entities were of the form cardinal, quantity and money.

([Python Notebook: https://colab.research.google.com/drive/1v-kInA9cW\_hZj--iSH8bnmv\_yKdhUdE0?usp=sharing](Python%20Notebook:%20https:/colab.research.google.com/drive/1v-kInA9cW_hZj--iSH8bnmv_yKdhUdE0?usp=sharing))

References:

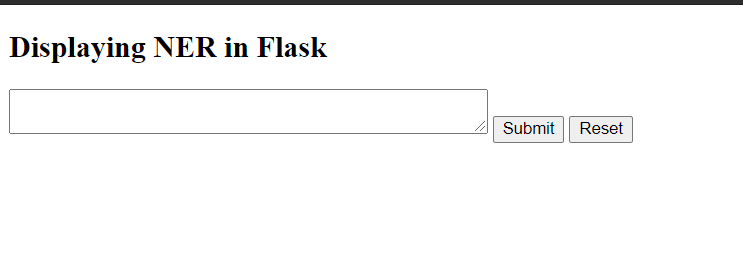
1. <https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1184/reports/6866023.pdf>

2. <https://spacy.io/>

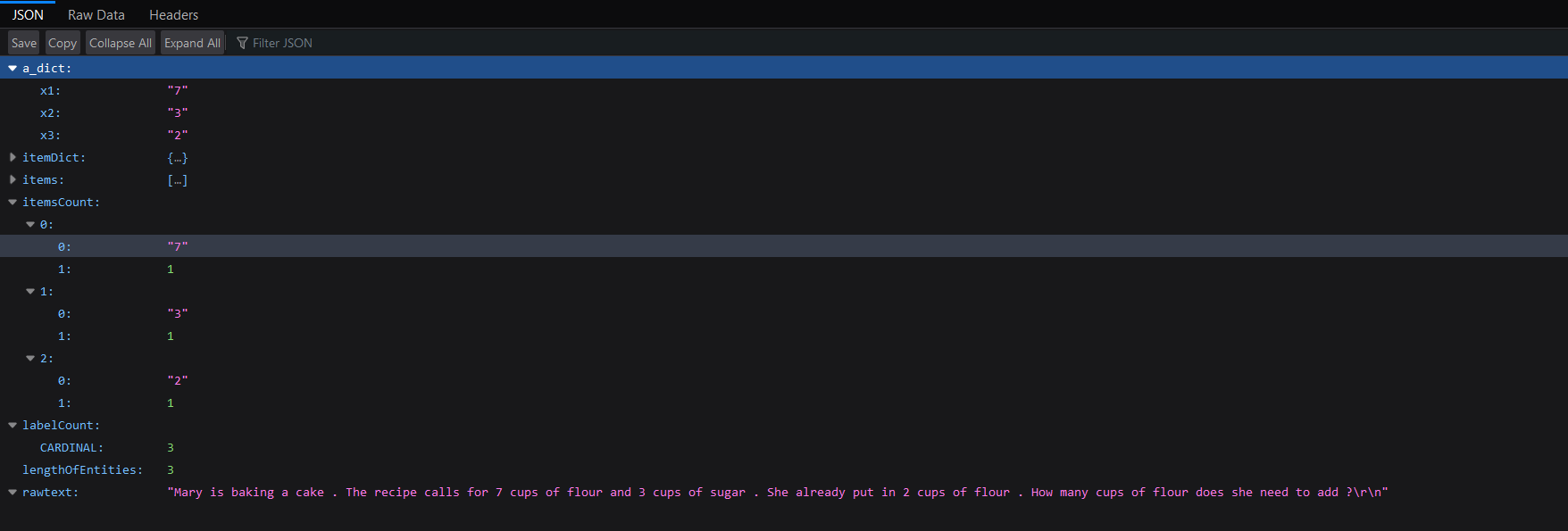
3. <https://www.nltk.org/>

How to use the code:

Go to <https://vishalwordproblem.herokuapp.com/>



Fill in the form with a word problem and click “Submit”



You’ll get the JSON response.

**Using CURL command:**

Curl -X POST -F 'rawtext= <The word problem>' https://vishalwordproblem.herokuapp.com/values